Hacia un método computacional para detectar léxico contrastivo

Damián Eliel Aleman

13 de noviembre de 2017

Temario

Qué es una palabra contrastiva

Se dice que una palabra es *contrastiva* cuando la frecuencia de uso en distintas regiones es muy diferente.

Ejemplos palabras contrastivas Argentina - España

- "che"
- "metegol"

Ejemplos palabras contrastivas dentro de Argentina

- "gurisada"

¿Para qué sirve conocer las palabras contrastivas?

Motivación de un método computacional para detectar léxico contrastivo

¿Cómo se conocían las palabras contrastivas? Mediante encuestas.

- Es costoso de realizar
- Muy difícil de hacer de forma balanceada en distintas regiones de un país o de un continente.
- Más difícil es encuestar a una gran cantidad de personas.
- Se basan en el conocimiento *a priori*

Temario

Web como un corpus

Kilgariff comentaba la riqueza de la Web para acceder a una información, antes impensada. Principalmente se hacían estudios estimando la frecuencia de una palabra viendo la cantidad de resultados de las consultas en un motor de búsqueda.

Ventajas:

- Gratuito
- Gran cantidad de datos
- Disponible e inmediato

Utilidades:

- Corrección de ortografía
- Traducción de frases
- Estimar el tamaño de la Web

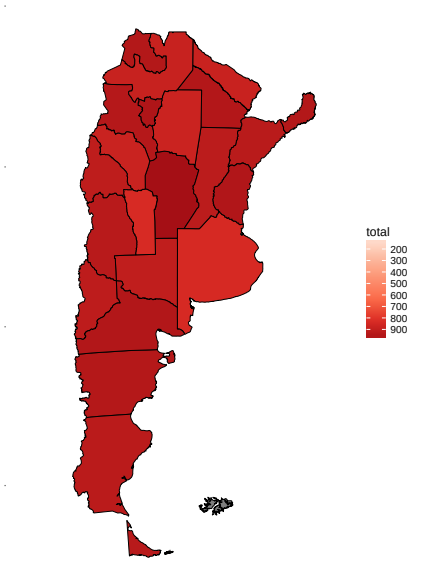
Cita

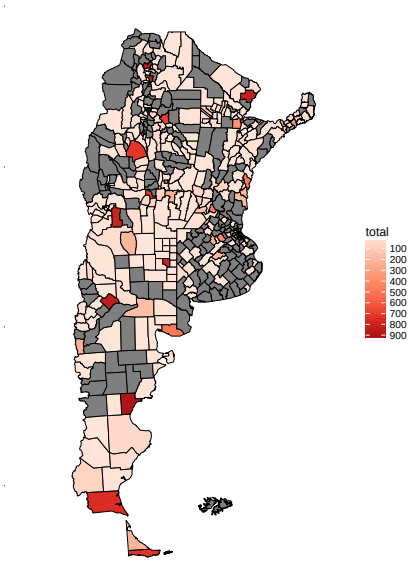
"La web es un corpus sucio, pero el uso esperado es mucho más frecuente que lo que puede considerarse como ruido." [KG03, p. 342]

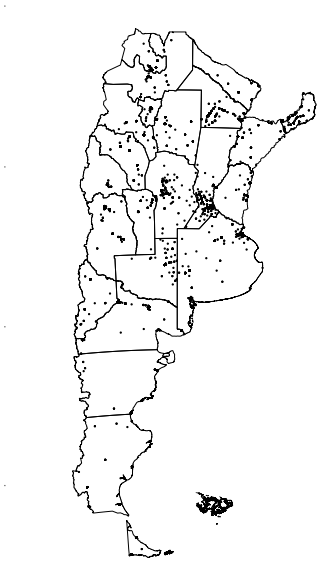
Temario

Búsquedas geolocalizadas

La búsqueda geolocalizada es una herramienta que nos da la posibilidad de obtener tuits generados en un área geográfica particular. Para esto, primero intenta buscar tuits cuyas coordenadas sean las buscadas.







Se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas.
Nos quedamos con los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia.

Tokenización y normalización del texto

Tokenización

Se consideró una palabra a las secuencias de caracteres formados únicamente por letras. Por lo tanto se eliminaron las menciones con @, los hashtags y links entre otros. Decidimos ignorar estos términos ya que no tienen interés lingüístico y agregarían mucho ruido a los datos.

Normalización

Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se hizo con la librería *TweetTokenizer de NLTK*.

Temario

Primeras métricas: MaxDif

Para cada palabra ω y cada par de provincias p_1 y p_2, podemos calcular el cociente entre la frecuencia máxima de ω en ambas provincias y la frecuencia mínima:

$$maxDif\,(\omega, p_1, p_2) = \frac{f_{max}(\omega, p_1, p_2)}{f_{min}(\omega, p_1, p_2)} \tag{1}$$

Desventajas:

1. Un valor para cada par de provincias.
2. No se considera la dispersión de los valores en todas las provincias.

Primeras métricas: $MaxDif_g$

Considerando las frecuencias de una palabra ω sobre todas las provincias, definimos:

$$maxDif_g(\omega) = \frac{f'_{max}(\omega)}{f'_{min}(\omega)} \tag{2}$$

donde $f'_{max}(\omega)$ es la frecuencia máxima de la palabra ω entre las frecuencias de todas las provincias y $f'_{min}(\omega)$ es la frecuencia mínima distinta de 0.
Con $maxDif_g$ se resume en un único valor la contrastividad de la palabra, sin embargo sigue sin considerar la distribución de las frecuencias.

La entropía de la información

La entropía nos brinda un valor que indica qué tan uniforme es la distribución de las frecuencias de cada palabra.
Las palabras con menor probabilidad son las que aportan más información.

Intuición

Las palabras más utilizadas como *de* o *que* aportan menos información que la palabra *celular*.

Observaciones

- La entropía es máxima cuando los eventos de X son equiprobables.
- La entropía es 0 si y solo si todas las probabilidades son 0 a excepción de una con probabilidad igual a la unidad.

Valor de la información

Zanette y Montemurro definieron al *valor de la información de una palabra* como

$$\Delta I_w(\omega) = p(\omega)\,(\widehat{H}(\omega) - H(\omega)) = p(w)\,\Delta H(\omega) \tag{3}$$

siendo $p(\omega)$ la frecuencia total de la palabra en el texto.

Valor contrastivo sobre las palabras

Valor contrastivo sobre las palabras

$$I_w(\omega) = norm_w(\omega) \cdot (\widehat{H}_w(\omega) - H_w(\omega)) \tag{4}$$

donde $norm_w$ sirve para normalizar sobre la cantidad de ocurrencias de la palabra.

Observaciones

- Si una palabra se dice muchas veces el valor de $I_w(\omega)$ es más alto.
- Si dos palabras se dicen la misma cantidad de veces, la palabra que tenga una dispersión más heterogenea será la de mayor valor contrastivo sobre las palabras.

Agregando la cantidad de personas que mencionan cada palabra

Hay palabras que pueden tener una frecuencia alta debido a pocas personas que las mencionan constantemente. Por eso se decidió también tener en cuenta a la distribución de la cantidad de personas que mencionan cada palabra.

Valor contrastivo sobre las personas

$$I_p(\omega) = norm_p(\omega) \cdot (\widehat{H}_p(\omega) - H_p(\omega)) \tag{5}$$

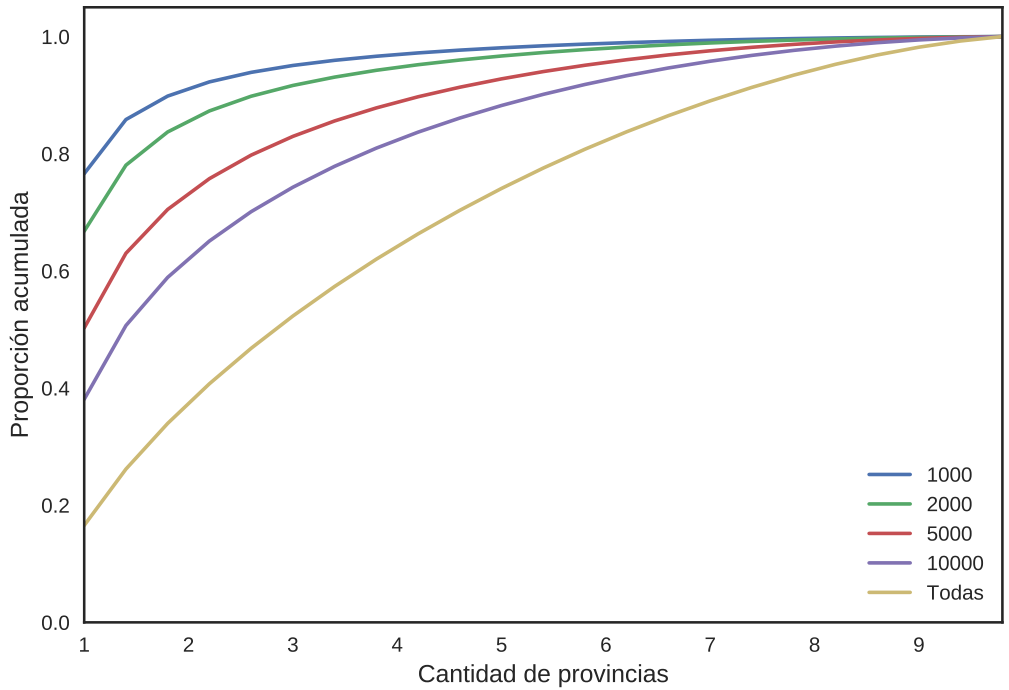
donde $norm_p$ sirve para normalizar sobre la cantidad de personas que mencionan la palabra.

Valor de contrastividad

Valor de contrastividad

Como nos interesa tanto la distribución de la cantidad de ocurrencias de cada palabra, como la de la cantidad de usuarios que la menciona, definimos

$$I(\omega) = I_w(\omega) \cdot I_p(\omega) \tag{6}$$

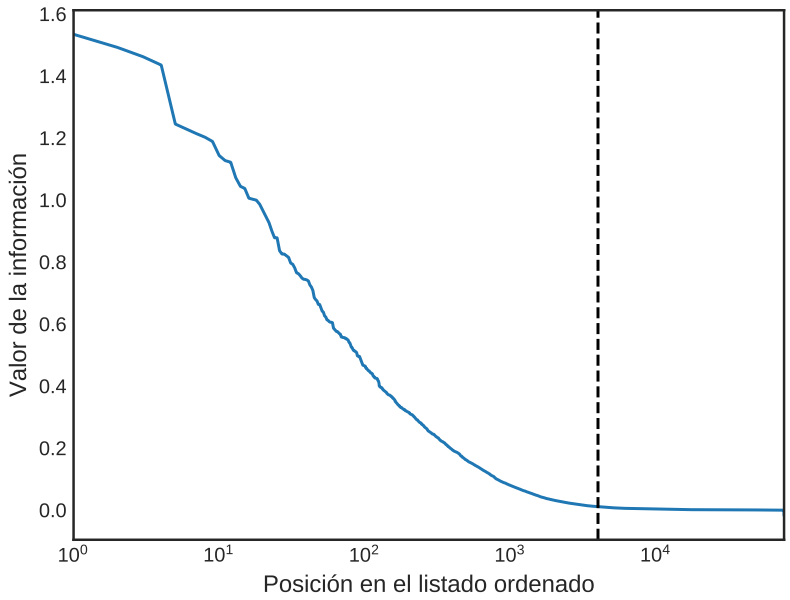


Proporción de ocurrencias acumulada según la muestra de palabras.

Temario

Palabras candidatas

Para buscar las palabras candidatas a tener contrastes significativos en cuanto a la cantidad de ocurrencias en distintas provincias, elegimos el conjunto de las primeras cinco mil (5000) palabras con mayor valor de nuestra métrica.



Validación lingüística

- **Coloquialismos o vulgarismos**
  "Perdon pero tenes que ser muy **culiado/a** para ir a mc y pedirte una ensalada" (Córdoba)
  "Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda" (Mendoza)
- **Indigenismos**
  "Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo (Del guaraní mitaí "pequeño")" (Formosa)
  "**Angá** mi negrito, esta triste (Del guaraní angá aprox. "pobre") (Corrientes)" (Corrientes)
- **Gentilicios**
  Casildense (de Casilda), **concordiense** (de Concordia) y **obereño** (de Oberá).

Test hipergeométrico

Para aplicar el test hipergeométrico representamos los datos sobre la palabra en una tabla de 2x2 como la de la siguiente Tabla.

	#Palabras sobre región	#Palabras en el resto de Argentina	Total
# Palabras w	k	$K - k$	K
# Palabras $\neq$ w	$n - k$	$N + k - n - K$	$N - K$
Total	n	$N - n$	N

Test t de Welch

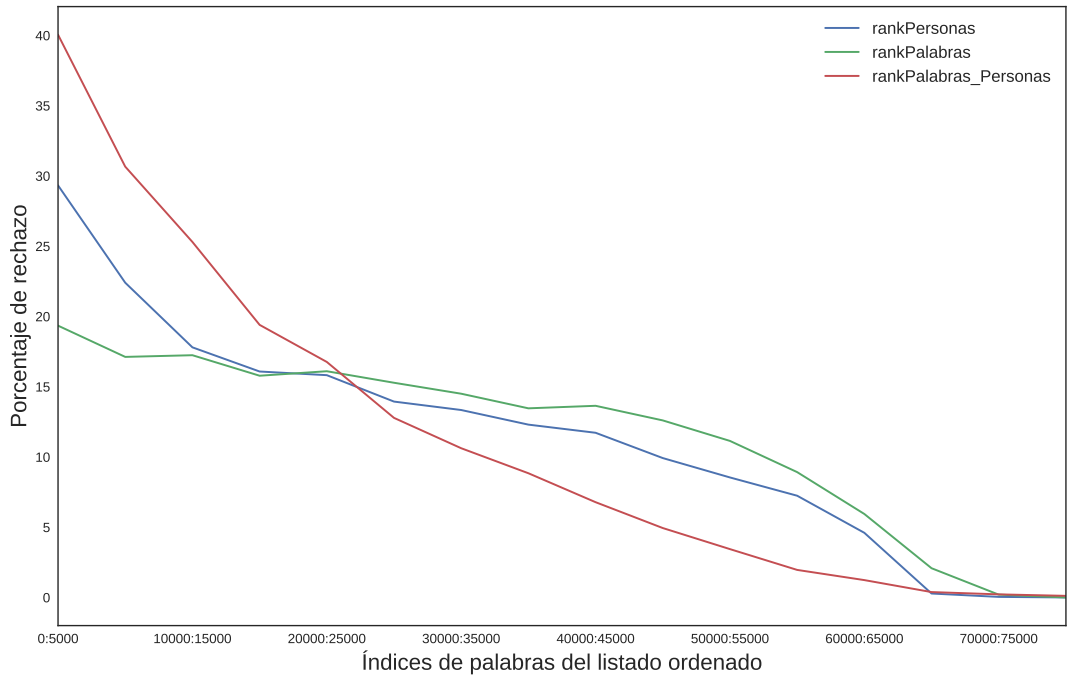
- El test de Welch nos provee un valor de probabilidad para rechazar la hipótesis nula que afirma que las medias de las dos distribuciones son iguales.
- Las suposiciones del test
  1. todos los textos son estadísticamente independientes
  2. la media de las frecuencias proviene de una distribución normal
- Agrupamos todos los tuits de cada usuario representando un texto.

  Corpus S todos los textos de los usuarios que provienen de las provincias en donde se cubre el 80 % de las ocurrencias

  Corpus T los textos creados por usuarios del resto de las provincias
- Notar que la suposición de independencia es más débil.

Resultados test t de Welch

La tasa de rechazo de la hipótesis nula en los distintos conjuntos de palabras, los cuales varían según el índice de estas en el listado ordenado según la métrica elegida.



Temario

Conclusiones

- Desarrollamos una métrica de la contrastividad de uso de una palabra en distintas regiones.
- Para probar esta métrica recolectamos un conjunto de datos de textos de la Argentina a través de la API de Twitter.
- Obtuvimos aproximadamente 1 palabra contrastiva relevante lingüísticamente cada 17 palabras.
- Varias de las palabras detectadas a partir de la métrica desarrollada serán agregadas al Diccionario del habla de los argentinos.

Trabajo a futuro

- Reproducir el trabajo para todos los países hispanoparlantes.
- Obtener regiones dialectales a partir de métodos de clustering, lo cual permitiría validar la vigencia de las regiones propuestas por Vidal de Battini en 1964 [VdB64].
- Analizar la contrastividad léxica comparando la distribución de n-gramas.

¿Preguntas?

Adam Kilgarriff and Gregory Grefenstette.
Introduction to the special issue on the web as corpus.
Computational linguistics, 29(3):333–347, 2003.

Berta Elena Vidal de Battini.
El español en la argentina.
Technical report, Argentina., 1964.